education sciences

Article

The Effect of Self-Monitoring on Mental Effort and Problem-Solving Performance: A Mixed-Methods Study

Madison Graham [1,2,*], Marinela Ilic [1,3], Martine Baars [4], Kim Ouwehand [1] and Fred Paas [1,5]

[1] Department of Psychology, Education, and Child Studies, Erasmus University Rotterdam, 3062 PA Rotterdam, The Netherlands; ouwehand@essb.eur.nl (K.O.); paas@essb.eur.nl (F.P.)
[2] Avans University of Applied Sciences, 4817 LL Breda, The Netherlands
[3] CED-Groep–Educational Advice and Development, 3067 GH Rotterdam, The Netherlands
[4] Behavioural Science Institute, Radboud University, 6525 XZ Nijmegen, The Netherlands; martine.baars@ru.nl
[5] School of Education, University of New South Wales, Sydney, NSW 2033, Australia
* Correspondence: graham@essb.eur.nl

Abstract: Self-regulated learning (SRL) has become increasingly important for learners in the 21st century as they engage more frequently in student-centered education. Accurate self-monitoring of one's learning process is central to SRL. However, research has shown that learners often struggle with monitoring accuracy. One possible explanation is that the process of self-monitoring itself may add cognitive load, potentially reducing both self-monitoring accuracy and overall performance. In the current study, we investigated the effects of self-monitoring on perceived mental effort and performance during problem-solving tasks of increasing complexity. Participants in the experimental condition were instructed to think aloud while problem-solving, while those in the control condition did not receive such instructions. Our findings indicate no interaction effect between the requirement for explicit and continuous self-monitoring and task complexity on perceived mental effort and performance. However, task complexity significantly impacted the perceived mental effort and performance in both conditions, while it affected monitoring accuracy exclusively in the experimental condition. Furthermore, qualitative analyses showed that the participants in the experimental conditions engaged in self-monitoring during 21.2 % of their think-aloud protocols. Contrary to previous research, self-monitoring did not appear to increase cognitive load, although the experimental participants took significantly longer to complete all the tasks. Possible explanations for these findings are discussed.

Keywords: self-regulated learning; cognitive load; self-monitoring; problem-solving; think-aloud protocols

check for updates

Citation: Graham, M.; Ilic, M.; Baars, M.; Ouwehand, K.; Paas, F. The Effect of Self-Monitoring on Mental Effort and Problem-Solving Performance: A Mixed-Methods Study. *Educ. Sci.* **2024**, *14*, 1167. https://doi.org/10.3390/educsci14111167

Academic Editor: Myint Swe Khine

Received: 31 August 2024
Revised: 19 October 2024
Accepted: 21 October 2024
Published: 26 October 2024

1. Introduction

With the advent of the 21st century, education has shifted paradigms to become increasingly student-centered, requiring the development of a new set of student skills that fit this educational model. One of these necessary student skills is self-regulated learning (SRL) [1]. Zimmerman [2] defined self-regulated learning as students taking charge of their learning through their (meta)cognitive processes, motivation, and behaviors. The need for SRL is evident in educational innovations such as problem-based learning [3], flipped classrooms [4], and massive open online courses (MOOCs) [5], as students in higher education are increasingly required to take charge of and critically evaluate their learning processes. This study focuses on metacognitive monitoring—a key component through which SRL functions [1,6].

However, learners often do not effectively monitor their own learning [7–11]. This may be because learning tasks that typically require good SRL skills such as self-monitoring are often high in intrinsic cognitive load, which could lead to a cognitive overload [11,12]. Cognitive load theory (CLT) provides an explanation for this since a central aim of CLT is to optimize instructional designs to fit the limited working memory capacity of the learner in a way that encourages better performance [13,14]. In short, the working memory is the workbench of the human information processing system and plays a key role in what information will be selected for consolidation in long term memory (e.g., [15]). Tasks that impose high cognitive load are challenging for this limited capacity working memory to process [16]. Furthermore, CLT distinguishes between three types of cognitive load: intrinsic load, which is inherent to the complexity of the task [17]; extraneous load, which arises from the learning environment (instruction, physical environment); and germane load, which refers to the cognitive resources allocated by working memory to manage the intrinsic cognitive [13,14,18]. The three types of cognitive load together make up the total cognitive load a learner has the handle during learning. Hence, next to the intrinsic load already imposed by the task itself, self-monitoring learning might heighten the total amount of cognitive load, leading to an overload for learners.

Research has shown that monitoring may be seen as a concurrent secondary task during a primary (learning) task [11,12], imposing additional cognitive load to the learning process. Therefore, learning may be affected negatively in such situations that already induce high cognitive load. Consequently, reducing the ability to accurately self-monitor one's own performance may be one of these undesirable effects during high cognitive load. For example, in a study by Radović et al. [19], it was found that item difficulty hindered the ability to self-monitor accurately. Therefore, the aim of the current study is to investigate if self-monitoring indeed adds to the cognitive load experienced by learners and thereby reduces monitoring accuracy and performance. Ultimately, this study investigated the effect of self-monitoring on perceived mental effort and performance with increasing task complexity during problem-solving tasks, as well as the effect of task complexity on monitoring accuracy.

Metacognitive Monitoring within Self-Regulated Learning

Self-regulated learning (SRL) is directed by the learners themselves, and as Pintrich [20] aptly wrote, it is "an active, constructive process whereby learners set goals for their learning and then attempt to monitor, regulate, and control their cognition, motivation, and behavior, guided and constrained by their goals and the contextual features in the environment" (p. 453). SRL is a recursive process of learners defining the task at hand, planning how to achieve their task goals, utilizing various study strategies, and adapting their studying using metacognition [21]. Of note are the metacognitive aspects of SRL, which are included in most models of SRL [22]. In essence, metacognition is thinking about one's own cognitions [23] and involves learners being aware and in control of their cognitions [24]. Metacognition can be further distinguished into different regulatory skills, such as planning, organizing, controlling, monitoring, and evaluating [2]. Of interest to the present study is monitoring as it compromises effective checking of understanding of information in tasks such as problem-solving [25]. Monitoring may be seen as a necessary pre-requirement for SRL [1], as learners must identify learning issues and reflect on how these issues are related to their learning process to remedy them [26].

Importantly, for an effective and efficient SRL cycle, monitoring must be accurate, meaning it should align closely with the learner's actual performance [1,27]. One measure of monitoring accuracy is calibration, which traditionally "refers to the accuracy of learners' perceptions of their own performance" [28] (p. 4). For this experiment, calibration will be referred to as self-assessment (SA) accuracy. Various studies have noted that learners frequently struggle with accurate monitoring [7,11], which may have detrimental effects on task performance. This issue arises because inaccurate monitoring leads to metacognitive control decisions based on incorrect judgments. Metacognitive control involves learners' decisions about implementing actions that would benefit their learning [29]. Effective metacognition requires frequent monitoring and control, which function interdependently [30]. Without adequate monitoring, there are no accurate evaluations of performance and effort and therefore no proper control actions taken to improve learning.

For a successful SRL cycle, it is essential that metacognitive control is grounded in accurate monitoring judgments.

Monitoring and Cognitive Load Theory

The central question of this study is whether self-monitoring increases cognitive load, under what conditions this occurs, and how it affects the learning process. While Seufert [31] argues that all phases of self-regulation, including monitoring, increase cognitive load, Peng and Tullis [29] found that monitoring may require only minimal cognitive resources. Given this discrepancy, cognitive load theory is particularly relevant to the study of SRL, especially concerning the monitoring process.

Relevant to the measurement of cognitive load for this study, Paas [32] differentiated cognitive load into two components: mental load, defined by the structural characteristics of a task, and mental effort, which represents the number of cognitive resources a learner uses to engage with the task. Importantly, Paas [32] highlighted that subjective ratings of mental effort needed to solve a task can serve as an indicator of cognitive load. In this experiment, self-reported mental effort ratings were used to measure cognitive load for each task. Additionally, Paas and Van Merriënboer [16] explained that the complexity of a task, or its mental load, often leads to increased mental effort. As task complexity increases, the intrinsic cognitive load experienced by the learner also rises. This can inhibit performance and learning, given the limited cognitive processing available to humans [16]. When intrinsic load increases in complex tasks, additional demands such as monitoring may contribute to cognitive overload, reducing the resources available for effective task processing. Consequently, this may to lead to inaccurate self-monitoring, impaired task performance, and hindered learning outcomes.

Optimizing the cognitive load experienced by the learner is crucial for achieving the most effective learning outcomes. However, since monitoring can consume cognitive resources [11], it is also important to optimize the amount of monitoring. As Winne [33] noted, monitoring that is excessive or non-directed may lead to undesirable effects. While prompts require only periodic monitoring, one study by Vangsness and Young [34] found that frequently prompting learners during a standardized exam resulted in worse performance. While Peng and Tullis's [29] study showed that monitoring may only need few cognitive resources, Brünken et al. [17] importantly noted that even if a secondary task needs few mental resources, it could still affect primary task learning. This may be because of the cost of task switching or divided attention. Frequently being prompted to monitor their learning, such as in Vangsness and Young [34], may then have required learners to switch tasks or multitask often.

Task switching has been shown to decrease working memory functioning [35], suggesting that task switching may cost cognitive resources. Additionally, high levels of multitasking during studying have been linked to lower exam performance than those who used less concurrent media technologies during studying [36]. Task switching from a primary task to a secondary task may induce a cost of additional cognitive resources that affects the performance of the primary task. In addition, divided attention may have other repercussions for monitoring accuracy and performance, as Peng and Tullis [29] found that it significantly hindered monitoring calibration accuracy and metacognitive control. Additionally, while people do see there are costs to multitasking, they are not good at predicting how much these costs affect their performance in memory tasks [37]. In this case, the secondary task of constantly monitoring learning may induce costs to the primary task by increasing mental effort, which then may cause cognitive overload as complexity of the task increases, possibly leading to decreased performance of the primary task.

Indeed, in Van Gog et al. [11], it was found that a self-monitoring instruction led to significantly higher cognitive load and lowered performance on a complex puzzle task. However, as Van Gog et al.'s [11] participants only monitored mentally, it is unclear to what extent and how often participants monitored their performance during the puzzle task. Hence, it remains unclear if and how self-monitoring adds to cognitive load and thereby could possibly hinder performance and learning.

Frequency of Monitoring

Another important question is how often do learners spontaneously monitor their learning? It is important to note that the frequency of monitoring may change from task to task, as some tasks may not require frequent monitoring to achieve proper understanding. For example, Trabasso and Magliano [38] showed that learners who read informational texts verbalized more metacognitive statements than those who read narrative stories. However, it is still an open question at what frequency learners would monitor their learning process when learning to solve problems accompanied by a prompt to continuously monitor this process.

The overall monitoring frequency when conducting a task may vary on many factors, but there have been several studies that looked at frequency of monitoring verbalizations using think-aloud protocols. In a study conducted with primary school students, it was found that they only verbalized monitoring for about 8.5 percent on average of their total verbalizations, but the participant percentage of monitoring fluctuated from zero to 27 percent [39]. The 20 teenage participants of Schellings et al.'s [40] study verbalized monitoring cumulatively for 30.6 percent of all the think-aloud protocols, which was second only to executing verbalizations at 39.9 percent. Though the participants were children in both studies, the ability to self-regulate and specifically monitor may increase as children grow older [41], indicating that adults would potentially verbalize monitoring more often than young children but potentially only slightly more than teenagers. However, it is still an open question as to what the monitoring frequency of university students during problem-solving would be.

Taken together, self-monitoring is a central aspect of SRL but could cause additional cognitive load which could harm the accuracy of self-monitoring and subsequent performance. Moreover, it is still unknown what the frequency of self-monitoring during learning to solve problems would be when students are instructed to explicitly monitor their learning aloud.

The Present Study

Think-aloud protocols are considered the most accurate representation of a participant's short-term memory [42]. Additionally, Sweller et al. [13] explained that people are only able to monitor information in the working memory. Therefore, the think-aloud sessions of the experimental group were recorded to ensure participants verbalized their thoughts consistently and to facilitate subsequent qualitative analysis. This approach also allows the experiment to explore the frequency of monitoring as captured in the think-aloud protocols.

From the literature described above, we drew the following hypotheses.

H1a. *There is a main effect of task complexity on mental effort, reflected in higher mental effort for the complex compared to the simple problems.*

H1b. *There is a main effect of the monitoring condition on mental effort, reflected in higher mental effort in the thinking-aloud condition compared to the control condition.*

H1c. *There is an interaction effect between the condition (thinking aloud and control condition) and task complexity on perceived mental effort. More specifically for complex problems, mental effort is expected to be higher in the thinking-aloud group than the control group, but no effect of the condition is expected for the simple problems.*

H2a. *There is a main effect of task complexity on performance, reflected in higher performance scores on the simple problems than on the complex problems.*

H2b. *There is a main effect of the monitoring condition on performance, reflected in higher performance scores in the thinking-aloud condition compared to the control condition.*

H2c. *There is an interaction effect between condition and task complexity on performance. More specifically, the performance scores of complex problems for the thinking-aloud group are expected to be lower than for the control group, but for the simple problems, there is no such expected effect.*

H3. *Within the thinking-aloud condition, there is a main effect of task complexity on self-assessment (SA) accuracy, reflected in lower SA accuracy for complex problems than for simple problems.*

2. Materials and Methods

Participants

Participants were recruited through the Erasmus Behavioral Lab Research Administration System (ERAS) from a population of undergraduate psychology students at Erasmus University, Rotterdam. In the information section of the study, the study was stated to be conducted in English and that all genders were being recruited. Participant recruitment started in the beginning of April 2023, and data collection lasted from the beginning of May 2023 through the beginning of June 2023.

A total of 65 participants participated in the experiment, with $n = 30$ in the experimental condition and $n = 35$ in the control condition. Originally, 70 participated in the experiment, but five were excluded due to technical difficulties or because they did not follow the instructions in the experiment. In the experiment, 51 participants identified as female, 12 identified as male, and 2 identified their gender as "Other". The mean age was 20.09 ($SD = 1.85$) years old, with participants all between the ages of 18 to 25. Only 16.9 percent of the participants reported that their (primary) native language was English.

Participants were randomized to each of the two conditions by way of alternating participants to each group. Through the computer command script used in the lab, subjects with an odd participant number were assigned to the experimental condition and subjects with an even participant number were assigned to the control condition. Participant numbering started at one and went up in numerical order until all participants had been tested.

All participants gave informed consent about the nature of the study and their rights beforehand and were instructed they could rescind their participation at any time during the experiment. Compensation for successful completion of the study was provided as undergraduate research credits.

Materials

Apparatus and Software

This study was conducted in the Behavioral Laboratory spaces of the hosting institute. Each participant completed the experiment in a closed and quiet individual cubicle with a computer and keyboard with a webcam with built in audio recording. The survey part of the experiment utilized Qualtrics [43], with one survey for the control and a separate survey for the experimental condition.

Demographics

The demographic section of the study asked three simple questions of a participant's age, gender, and native language. Gender options were listed as female, male, or other. The native language question could be answered by filling out a text box. If a participant put more than one native language in the text box, only the first one written was included in determining the amount of native English speakers.

Pretest

The pretest consisted of nine open-ended questions, adapted from Baars, Vink et al. [10], to assess each participant's theoretical knowledge of heredity. Participants had 10 minutes to answer all nine questions until the pretest was automatically submitted. They were instructed to answer each question in a sentence or two in English, and if they absolutely could not answer a question, to put a question mark as the answer. An example pretest question is: "What is the genotype of a hereditary quality such as hair color?" The full list of pretest questions can be found in Appendix A.

The pretest performance was assessed by the researchers through a keyword grading scheme that was made by the researchers before the data collection started. The keyword grading scheme used can be found in Appendix B. If the researcher saw the required keywords were present in the answers, the participant received a point per question for nine possible points. If participants did not include the proper keywords but could still describe the answer to the question well enough, the answer was marked as correct. An example of this can also be seen in Appendix B.

Learning Phase

Brief information on genotypes, homozygous and heterozygous alleles, cross tables, and the types of reasoning needed to solve the problems was presented to all participants before the main problem-solving tasks. The cross-table information was not originally included in this learning phase material from a previous master thesis who used it, but it was added after testing the procedure with some undergraduate research assistants who gave feedback that it would be beneficial to include. The cross-table information was added by the author, agreed upon by both primary researchers, and tested again with another undergraduate research assistant. The exact information provided in the whole learning phase can be found in Appendix C.

Instructions for Problem-Solving Tasks

Participants in the control group were instructed they would solve 10 problems. They were instructed to solve the problems as quickly and accurately as they could and if they could not answer a question to fill in a question mark. They were not allowed to use their hands, write things down on paper, or speak when solving the problems. Participants in the experimental group received similar instructions as the control group but were additionally provided with written instructions to think aloud (based on Moos and Bonde [44]) in English: "Please say what you're doing or thinking while you are doing or thinking it". While their audio was recorded, participants were also instructed not to give any personally identifiable information during the recording, not to write anything down on paper, and not to use their hands. It was instructed that the audio recording would start automatically, so they were to start speaking as soon they began to solve the problems. The exact instructions provided to the experimental group can also be found in Appendix G.

Heredity Problem-Solving Task

In this phase, participants were asked to quickly work through ten problem-solving tasks about Mendelian inheritance. The participants solved the tasks in the learning environment where they could write down their answers per step. Both conditions received the same set of tasks. There was no time limit imposed per problem-solving task. The tasks were adapted from material from Baars, Vink et al. [10]. The basic steps of each question, as described by Baars, Vink et al. [10], included determining possible genotypes from phenotypes, completing a heredity diagram, establishing a direction of logical reasoning, calculating how many cross tables were needed, filling in the cross tables, and deriving the final answer from them. There were five complexity levels in total with two questions per complexity level, presented in ascending order of complexity. Three examples of tasks from this experiment can be found in Appendix D. The higher complexity levels had more steps per question than the lower complexity levels. Level four had an extra direction of reasoning and cross table section than previous levels, and level five had an additional final answer multiple choice step on top of that. For the analyses, simple tasks were defined as complexity levels one and two, whereas complex tasks were defined as complexity levels three through five (see Appendix J for the rationale behind this decision).

Performance on the problem-solving tasks was graded by the researchers with an answer sheet that was made by the researchers before the data collection started. In each task, only the steps of determining genotypes, the heredity diagram, the direction of reasoning, and the steps with the final answers were graded. Participants received one point per correct step. For the first three complexity levels, participants could receive a maximum of four points per problem-solving task, for complexity level four they could receive a maximum of five points, and for complexity level five they could receive a total of six points maximum. Mean percentage scores were calculated for the simple (level 1 and 2) and difficult problems (level 3, 4 and 5) by dividing the score of the participant by the perfect score and averaging them across the simple and complex levels, respectively.

Invested Mental Effort Rating Scale

The participant's subjective mental effort was measured with a single-item nine-point scale, in which participants rated how much mental effort they invested to solve the just previously shown problem-solving task, with a (*1*) *very, very low mental effort* rating ranging to a (*9*) *very, very high mental effort* rating [32]. The reliability of the scale was estimated with Cronbach's coefficient alpha. A coefficient of reliability (α) of 0.90 was obtained in the study of Paas [32]. An example of the scale used in the survey can be found in Appendix E. Mean mental effort ratings were calculated for the simple and complex problems.

Absolute Self-Assessment (SA) Accuracy Measures

A self-assessment rating of their own performance was only given to participants in the experimental condition, so as to not have the control condition unintentionally monitor their learning. For the first three complexity levels, participants rated how many points they thought they scored from *0 points* to *4 points*. For complexity level four, participants rated how many points they thought they scored from *0 points* to *5 points*. Finally, for complexity level five, they rated how many points they thought they scored from *0 points* to *6 points*. Examples of the scales can be found in Appendix F. After converting the task performance points and self-assessed points to percentages, absolute SA accuracy was determined by subtracting the self-assessed percentage from the actual performance percentage per task. Then, the absolute value of this accuracy percentage was taken to obtain the absolute SA accuracy, showing the deviation from a perfectly assessed score of zero. Additionally, for each complexity level a mean SA score was calculated. To obtain the mean SA score for the simple tasks, the absolute SA accuracies from complexity levels one and two were all averaged together, and the same was done for levels three through five to obtain the mean SA score for the complex tasks.

Procedure

At the beginning, participants were instructed to wait until the experiment appeared on the screen and to read everything very carefully and follow written instructions. If they had questions they could press a help button, and a researcher would come to answer their question during the experiment. Once finished, they could leave the cubicle, and the researcher would shut down the experiment for them. They completed the experiment alone in a sound-isolated cubicle.

Upon starting the experiment, participants gave informed consent and then reported some demographic information (see the participants and demographics sections for a detailed description). Next, participants completed the heredity pretest section to obtain a baseline assessment of their theoretical knowledge of heredity. Then, in the following learning phase, basic information on genotypes and types of reasoning needed to answer the heredity problem-solving tasks was provided. Afterwards, they moved onto the instructions of the heredity problem-solving task for their respective groups. Control participants did not receive instructions to think aloud, while the experimental participants did, as can be seen in Appendix G. Participants were then asked to answer each of the ten problem-solving tasks. Reminders to think aloud in bold red text were written on every task in the experimental survey version.

In between each of the ten problems, all participants rated their subjective mental effort spent on the previous problem with the mental effort rating scale. Additionally, after each question, those in the experimental condition completed the self-assessment rating. After a participant had completed all 10 problems with accompanying rating scales, the experiment ended, and the recording stopped for the experimental group. Because there were no time limits on the task, the total duration of the experiment ranged between 26 to 78 minutes.

Qualitative Analysis Approach

Audios were transcribed non-verbatim by two research assistants and the first two authors. Transcripts were then segmented by each step in the problem-solving tasks. An adapted version of the base coding scheme by Azevedo et al. [45] (see Appendix H) was used to code the transcripts. The first adaptation we applied was splitting the Judgement of Learning (JOL) code into positive and negative JOLs for more in-depth analysis of the kinds of JOLs the participants verbalized. This would lead also to a better understanding of how participants self-monitored. Secondly, due to the presence of many responses that could not be categorized with the base scheme, we added codes for verbalizations on the reading process (named "Reading the Question"), cognitive load ("Mental Effort), and coming to an answer ("Stating Answer"). Third, we added a Self-Assessment code for monitoring verbalizations when we prompted participants to self-assess in the task. Thirdly, we removed the Self-Questioning code of Azevedo et al. [45] due to redundancy with the Rereading code, as well as the code Selecting a New Informational Source and Control of Context as participants were unable to execute the definition of these codes in this study.

All codes were given new anchor examples from the transcripts if possible or given a researcher-derived example. Ten transcripts were selected out of the 30 via a random number generator (cf. Johnstone et al. [46]). Because in two transcripts most verbalizations were too unclear, we decided to exclude these from further analysis, leaving eight total transcripts included in the final qualitative analysis. To prepare for the intercoder reliability (ICR) analysis, about 30 percent of the total number of segments in the transcripts used in the analysis were selected (cf., Azevedo et al. [45]), which gave a selection of 17 segments per transcript that were chosen via a random number generator. The segments were compiled into one document and imported into Atlas.ti Version 24 [47] for the ICR analysis. Coder one freely coded the ICR document with the adapted coding scheme, then removed their codings but kept the quotations in the document. Coder two then independently coded the pre-chosen quotations. Cohen's kappa measures were calculated per code group in SPSS Version 29 [48]. Cohen's kappa was computed to determine if there was agreement between coder one and two on which code group each verbalization in the ICR analysis belonged to. There was fair agreement for code group "A — Forethought, Planning, and Activation", $\kappa = 0.401$, $p < 0.001$, 95% CI [0.187, 0.615]. There was substantial agreement for code group "B — Monitoring", $\kappa = 0.401$, $p < 0.001$, 95% CI [0.187, 0.615]. There was substantial agreement for code group "B—Monitoring", $\kappa = 0.685$, $p < 0.001$, 95% CI [0.587, 0.783]. For the code group "C — Strategy Use", there was moderate agreement, $\kappa = 0.498$, $p < 0.001$, 95% CI [0.408, 0.588]. For the code group "D — Task Difficulty, Complexity, and Demands", there was perfect agreement, $\kappa = 1.000$, $p < 0.001$, 95% CI [1.000, 1.000]. Lastly, for the code group "E — Miscellaneous Actions", there was moderate agreement, $\kappa = 0.543$, $p < 0.001$, 95% CI [0.449, 0.637]. Disagreements between coders were resolved via discussion to reach full agreement. Both coders gave their reasoning for a chosen code and changed their answer if they agreed with the other coder. Full agreement was important in this case because one coder would code all eight transcripts. Coder one then coded all eight transcripts using Atlas.ti [47]. Additionally, to obtain a frequency and percentage of Monitoring, the Self-Assessment codes were removed from the Monitoring code group.

3. Results

Pretest

A two-tailed independent sample *t*-test revealed no significant differences between the control and experimental conditions on total pretest scores: $t(63) = 1.76$, $p = 0.084$, $d = 0.44$. This indicates that both groups had similar performance on the pretest, suggesting no significant differences in their prior knowledge of Mendelian inheritance.

Quantitative Analysis Approach

Separate one-way mixed ANOVAs were conducted for mental effort and performance, with two factor levels with the within-subject variable for task complexity (simple vs. complex) and the between-subjects variable for condition (thinking aloud vs. no prompted monitoring), using a Bonferroni confidence interval adjustment. A one-way ANOVA was conducted on self-assessment (SA) accuracy (in the experimental condition only) with a within-subjects factor for task complexity (simple vs. complex), with a Bonferroni confidence interval adjustment.

All statistical analyses were conducted with a significance level of $\alpha \leq 0.05$ unless otherwise stated. Effect sizes were noted where appropriate with Cohen's [49] guidelines on the partial eta-squared effect size being small at 0.01, a medium effect at 0.06, and a large effect at 0.14. The quantitative data were analyzed with SPSS Version 29 [48].

Mental Effort Ratings

There was a main effect for task complexity, $F(1, 63) = 109.62$, at 0.01, a medium effect at 0.06, and a large effect at 0.14. The main effect of task complexity showed that participants rated higher mental effort for the complex problems compared to the simple problems (see Figure A1 of Appendix I). See Table 1 for the means and standard deviations for both simple and complex across conditions for all variables.

Table 1. Means and standard deviations of performance (range 0–100), mental effort ratings (range 1–9), and absolute value of self-assessment accuracy percentage (0 = perfect, 100 = largest deviation in monitoring) in the experimental and control conditions in simple and complex tasks.

		Control (n = 35)		Experimental (n = 30)	
		M	*SD*	*M*	*SD*
Simple	Mental Effort	3.96	2.08	3.83	1.57
	Performance	91.79	14.60	86.25	19.03
	SA Accuracy			13.33	16.39
Complex	Mental Effort	5.51	1.68	5.55	1.39
	Performance	86.29	19.79	76.13	24.05
	SA Accuracy			18.30	12.08

Note: There is no absolute self-assessment (SA) accuracy data for the control group as they did not self-assess after any of the problem-solving tasks. An absolute SA accuracy score of zero is a perfectly assessed score, meaning that there was no difference between the actual performance and the self-assessed score.

Performance Scores

There was a main effect for task complexity, $F(1, 63) = 20.77$, $p < 0.001$, ${\eta_p}^2 = 0.25$, but not for condition $F(1, 63) = 3.00$, $p = 0.088$, ${\eta_p}^2 = 0.045$. The interaction effect of task complexity and condition on performance was not significant, $F(1, 63) = 1.83$, $p = 0.182$, ${\eta_p}^2 = 0.028$.

The main effect of task complexity reflects the lower performance on the complex versus the simple problems (see Figure A2 of Appendix I).

Absolute Self-Assessment Accuracy in Experimental Group

There was a main effect of task complexity, $F(1, 29) = 5.62$, $p = 0.025$, ${\eta_p}^2 = 0.16$, reflecting higher SA accuracy for the complex problems than the simple problems (see Figure A3 of Appendix I). This suggests that participants in the experimental condition became significantly less accurate at evaluating their performance on complex tasks than on simple tasks.

Time to Complete Study

The experimental group on average ($M = 45.40$, $SD = 11.24$) took more time to complete the entire study than the control group ($M = 38.79$, $SD = 8.96$). According to a two-sided independent sample t-test, there was a significant difference between the two groups for the time to complete the study, $t(63) = -2.64$, $p = 0.011$, $d = -0.66$. This finding reflects that the participants in the experimental condition, in which they were instructed to explicitly monitor their learning process, invested more effort, in this case time, to learn how to solve the problem-solving tasks.

Qualitative Results on the Code Frequencies

Across all eight transcripts analyzed, there were 1536 coded verbalizations. For the Monitoring code group, there were 351 coded verbalizations, or about 22.9 percent of all codes as (see Table 2). For the frequency of Spontaneous Monitoring, which required removing the prompted Self-Assessment code frequencies within the Monitoring code group, there were then 325 coded verbalizations, or about 21.2 percent of all coded verbalizations. Additionally, because not every participant verbalized monitoring the same amount, an average frequency and percentage per transcript was calculated. For the Spontaneous Monitoring group specifically, this gave a mean frequency of 40.63 coded verbalizations per transcript, or about on average 21.2 percent of coded verbalizations per transcript. While more examples can be found in Appendix H, one example of a Negative JOL verbalization was "I don't know if that's good". Another example of a Positive JOL verbalization was "Which makes sense, because their other parent has blue eyes". Finally, an example of Content Evaluation verbalization from a transcript was "You can't guess from the information".

Table 2. Self-regulated learning code frequencies across all eight transcripts.

Codes	*Raw Frequency*	*Percentage*
A—Forethought, Planning, and Activation	***51***	***3.32***
A1—Planning	*7*	*0.46*
A2—Sub-Goals	*19*	*1.24*
A3—Prior Knowledge Activation	*12*	*0.78*
A4—Recycle Goal in Working Memory	*13*	*0.85*
B—Monitoring	***351***	***22.85***
B1—Negative Judgment of Learning (NJOL)	*41*	*2.67*
B2—Positive Judgment of Learning (PJOL)	*71*	*4.62*
B3—Self-assessment	*26*	*1.69*
B4—Feeling of Knowing (FOK)	*38*	*2.47*
B5—Content Evaluation	*115*	*7.49*
B6—Identify Adequacy of Information and Task Instructions	*60*	*3.91*
C—Strategy Use	***480***	***31.25***
C1—Goal-Free Search	*0*	*0*
C2—Summarization	*9*	*0.59*
C3—Rereading	*18*	*1.17*
C4—Inferences	*311*	*20.25*
C5—Hypothesizing	*25*	*1.63*
C6—Knowledge Elaboration	*117*	*7.62*
D—Task Difficulty, Complexity, and Demands	***43***	***2.80***
D1—Time and Effort Planning	*9*	*0.59*
D2—Help-Seeking Behavior	*1*	*0.07*
D3—Task Difficulty	*9*	*0.59*
D4—Mental Effort	*24*	*1.56*
E—Miscellaneous Actions	***611***	***39.78***
E1—Stating Answer	*72*	*4.69*
E2—Copying Information	*233*	*15.17*
E3—Reading Question	*306*	*19.92*

Note: Due to rounding, percentages within the code groups may not add up exactly to the total percentage of the code group.

4. Discussion

This study has contributed to the scientific understanding of cognitive load theory and SRL by investigating the effect of self-monitoring on the perceived mental effort and performance during problem-solving tasks of increasing complexity. In line with H1a and H2a, participants rated higher perceived mental effort when working on the complex compared to the simple problems and performed better on the simple than complex problems. Contrary to the expectations outlined in H1b and H2b, there was no significant effect of the monitoring condition on either mental effort or performance. In addition, we found no support for H1c and H2c, as task complexity had a similar effect on mental effort ratings and performance in both the experimental (monitoring via thinking aloud) and the control condition. In line with H3, participants showed more accurate self-assessments for the simple than the complex problem-solving tasks.

In sum, our findings indicate that there was no interaction effect between having to explicitly and continuously monitor and the task complexity on perceived mental effort or performance. However, participants in the experimental condition did spend more time on learning how to solve the problem-solving tasks compared to the participants in the control condition. In addition, the task complexity did affect the perceived mental effort and performance as well as the monitoring accuracy (only in the experimental condition). Furthermore, the qualitative analyses showed that the participants in the experimental conditions on average were engaged in self-monitoring in 21.2 % of their think-aloud protocols. While this percentage of monitoring in the think-aloud protocols is considerably lower than that which previous studies found for the self-monitoring of children [39,40], this type of heredity problem-solving task was not new to the participants as they had already encountered it in their undergraduate coursework. It is possible then that not as much self-monitoring for this specific task was needed, as they may have already been fluent at the task. However, this does indicate that the instruction that was given to elicit explicit monitoring worked. Nonetheless, monitoring did not seem to increase the cognitive load, in contrast to earlier research by Van Gog et al. [11]. There are some possible explanations for this difference that will be discussed below.

Importantly, the manipulation of the intrinsic cognitive load of the task was successful, as can be seen by the significant positive effect of task complexity on mental effort ratings. However, Paas [32] divided cognitive load into mental load and mental effort, stating that mental load was the instructional demands of a task versus mental effort being the invested mental effort to the task. Paas [32] denoted that this invested mental effort was reflective of the cognitive load of a task. Additionally, Leppink et al. [50] wrote that task complexity helps to determine intrinsic load but also the prior knowledge that participants have. Recent literature has even questioned the use of certain cognitive load-related appraisals like mental effort as they can be biased by other cues like response time [51]. It is not entirely clear, when using invested mental effort measures, if it is possible to obtain a pure measure of intrinsic cognitive load. However, because the format and wording of the steps of the ten increasingly complex problems were kept as consistent as possible with limited distractions, it is thought that extraneous cognitive load was kept at a constant. However, that leaves the possibility that the invested mental effort could also have been affected by intrinsic and germane load. It is possible that as intrinsic cognitive load increases, germane load increases as well for participants to handle the intrinsic cognitive load. Notably, Paas et al. [18] described that germane load can be considered as the resources needed by the working memory to manage the intrinsic load. Therefore, though the invested mental effort ratings did increase significantly as intended as the task complexity increased, it could have been caused by other factors next to intrinsic load, such as germane load.

There was also a significant negative effect of task complexity on performance, which aligns with H2a. As Paas and Van Merriënboer [16] noted, performance and learning during complex tasks can be hindered by the limited cognitive processing capacity of humans. As the intrinsic load increased, it likely placed greater demands on the participants' available cognitive resources while they engaged in the problem-solving tasks. On simple tasks, this may not have been an issue as the general cognitive load of the task was low, leaving plenty of resources to process the material. However, once the intrinsic cognitive load of a task became too high on the highest complexities, there may have been not enough cognitive resources for the germane load to adequately help the processing of the highest complexity level tasks. This overload may then have the consequences of incorrect and inefficient processing of the materials, possibly leading then to decreased performance among both groups due to the increasing complexity of the tasks. These results indicate that our working memory plays an important role in problem-solving, especially when a task becomes more complex. Teachers should be aware of this when checking the complexity of a task. If the task becomes too complex it may negatively impact the learning process and may lead to worse performance.

Notably, no significant interaction effects between self-monitoring and task complexity on mental effort or performance were observed, not giving evidence towards H1c nor H2c. While it was expected to find similar results to Van Gog et al. [11], this was not the case. However, a significant difference between the experimental and control condition in the time the participants spent on learning how to solve the problem-solving tasks was found. This indicates that the participants in the experimental condition, in which they explicitly monitored their learning process, did invest more effort (i.e., time) but did not reach a higher performance compared to the control condition. This possibly also explains the difference between the findings in the current study and the study by Van Gog et al. [11]. While Van Gog et al. [11] only used two Sudoku puzzle tasks, one simple and one complex, they gave their participants a time limit of two minutes per puzzle. Importantly, Paas and Van Merriënboer [16] noted that tasks with time limits often have a higher cognitive load. If there had been a time limit imposed per question in this study, it is hypothesized that there would have been an interaction effect of task complexity and monitoring condition. Setting a time limit per problem-solving task potentially would have caused higher cognitive load within the think-aloud condition, as then the think-aloud condition would not have the time and subsequent cognitive resources available to task switch between solving the task and constantly monitoring their learning by thinking aloud. This could possibly create even higher demands on participants' cognitive resources, which could be reflect in their perceived mental effort and performance. These results indicate that time is an important factor and should be taken into account when giving students a complex task. By giving students enough time to finish their work, students will not experience more mental effort while monitoring their learning.

Though there were frequent reminders to think aloud within the experimental version of the Qualtrics survey and audio was checked to make sure participants spoke consistently, participants did take pauses in their thinking aloud, often when executing an action such as typing. These pauses may have served as a miniature mental reprieve from monitoring, or they can be seen as microbreaks [52]. Previous studies like Tyler and Burns [53] have shown that 10-minute breaks between regulatory tasks can replenish self-regulatory resources, resulting in similar performance by experimental and control participants. Though there is no ideal micro-break duration established yet [52], another study has shown even just 3-minute breaks were beneficial for performance [54]. Without time pressure added, participants in the experimental condition in the present study could take brief pauses at their leisure, which may have added up, allowing them to potentially gain cognitive resources back over time. This may be a possible explanation for the similar mental effort and performance found in both conditions across the complexity levels, and for why the experimental condition took significantly longer to complete the whole experiment. Therefore, a lack of time limit per question may be one explanation for the results found. Future research could investigate the effect of self-monitoring on mental effort and performance while using a time limit per problem-solving task to investigate this further.

In addition, there was a significant effect of task complexity on absolute SA accuracy, meaning in the context of the data the experimental group participants were significantly worse at judging their performance in the complex tasks than in the simple ones. They had a greater absolute deviation between actual and self-assessed scores on the complex tasks. This is in line with results found in earlier research by Baars et al. [8,9] and Radović et al. [19], who found that SA accuracy decreased with increasing task complexity. Additionally, this is in line with previous research by Nietfeld et al. [55] which found that students showed more accurate judgements of performance on easier items than more difficult ones. Just as the increasing task complexity affected both mental effort and performance, it may be that, due to the cognitive overload on the highest complexity levels caused by the higher intrinsic and germane load, there were fewer cognitive resources available to accurately monitor their learning [12]. When complexity becomes too high, then it may have negative effects on learners' ability to judge their own performance.

There are some limitations to the current study. First, the lack of a time limit per problem-solving task made it harder to find an effect of self-monitoring on mental effort and performance. However, participants evidently needed more time to explicitly monitor their learning process, which could also be seen as an indication of an effect of self-monitoring on the effort participants invested in the learning process. Second, there was an inconsistent point system between the complexity levels in the problem-solving tasks. For example, in level one participants could score up to four points but for level five participants could score up to six. While the performance scores and SA accuracy were converted to percentages for the analysis, if this study were to be redone, a consistent point system could be arranged. Additionally, due to the open-ended nature of questions, all grading had to be done by the first two authors, potentially introducing bias when answers were partially correct. This could have been corrected by making the entire experiment able to be graded by computer. Finally, the small sample size may have impacted the statistical power and the reliability of our findings. While our limitations could have influenced the results, the observed differences in performance between groups, though not statistically significant, might have reached significance with a larger sample size or if a time limit had been imposed.

5. Conclusions

In the current study, the effect of self-monitoring using a think-aloud method on mental effort and performance was investigated for problems at increasing complexity levels. Although no interaction effect between self-monitoring and task complexity was found for mental effort and performance, participants in the experimental condition did spend more time on learning how to solve the problems. In addition, task complexity significantly increased mental effort and lowered performance in both conditions, and it lowered self-assessment accuracy (in the experimental condition only). These findings indicate that both monitoring and task complexity can take up cognitive resources from participants, which consequently affects their mental effort, self-assessment accuracy, performance, and the time they need to work on problem-solving tasks. As participants in the experimental condition used more time to work on the problem-solving tasks, it is possible that they were able to replenish their cognitive resources over time to be able to perform similarly to the control condition in both simple and complex tasks. Therefore, the time given to solve a task seems to be an important factor when considering self-monitoring. Additionally, it was found that participants in the experimental group verbalized self-monitoring for on average 21.2% of the think-aloud protocols, meaning that the experimental instructions elicited self-monitoring. Future research could investigate the effect of self-monitoring on mental effort and performance while using a time limit per problem-solving task to investigate this further. Implication-wise, because task complexity had an effect on self-assessment accuracy, teachers should be mindful of the complexity of a task when asking to students to self-assess.

Author Contributions: Conceptualization, M.G., M.I. and M.B.; methodology, M.G., M.I. and M.B.; data collection and analysis, M.G. and M.I.; writing—original draft preparation, M.G., M.I., M.B., K.O. and F.P.; writing—review and editing, M.G., M.I., M.B., K.O. and F.P.; supervision, M.B., K.O. and F.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board (or Ethics Committee) of Erasmus University Rotterdam (protocol code ETH2223-0393 and 8 March 2023).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Acknowledgments: The authors express gratitude to the Erasmus Behavioral Lab for their invaluable contributions to this study. The authors would also like to thank the three bachelor-level research assistants, Ivan Orekhov, Yu Kong, and David Hahn, who helped with testing the experimental procedures, data collection, and transcribing

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Hereditary Pretest

1. What is the genotype of a hereditary quality such as hair color?
2. What is the phenotype of a hereditary quality such as eye color?
3. What does it mean if a person is homozygous for a hereditary quality such as hair color?
4. What does it mean if a person is heterozygous for a hereditary quality such as eye color?
5. What does it mean if an allele in a genotype is dominant?
6. What does it mean if an allele in a genotype is recessive?
7. Which feature will be visible if the same recessive alleles are present in the genotype?
8. Which feature will be visible if two different alleles are present in the genotype?
9. What do you have to do to find the genotypes of children if you know the genotypes of the two parents?

Appendix B

Keyword Grading Scheme for Pretest

Question	Answer	Keywords
1	Genotype refers to the two alleles a person/organism has inherited for a particular gene	1. Alleles 2. Gene(s)
2	Phenotype is how this genetic information (genotype) is reflected in physical characteristics	1. Genetic information/genotype 2. Physical
3	Homozygous means that both alleles are the same	1. Alleles 2. both are the same/other words that describe this
4	Heterozygous means that there is dominant and a recessive allele	1. Alleles 2. Different
5	A dominant allele produces the dominant phenotype. Even with one dominant allele they will have the dominant phenotype	1. Dominant phenotype 2. one copy (3. Uppercase)
6	A recessive allele produces the recessive phenotype. To produce this phenotype the genotype must have two copies of the recessive allele.	1. Recessive phenotype 2. Two copies needed (3. Lowercase)
7	The recessive phenotype	1. Recessive
8	The dominant phenotype	2. Dominant
9	Put the genotypes of the parents in a cross table so you can form the possible combinations of genotypes for the child.	1. Cross table or Punnett Square

An example of not using the keywords but being given a point on the pretest:

Question one was "What is the genotype of a hereditary quality such as hair color?" One participant answered, "its the genetics passed on so like HH or sm". This was given a point for correctness because while they did not use either of the keywords of gene(s) or allele(s) for their answer, they did use the word genetics, and it was clear to the researchers that they understood that from Mendelian inheritance that a genotype was made up of two alleles which are represented with two letters, either upper- or lower-case.

Appendix C

Learning Phase Material

In the upcoming section, you will be asked to solve 10 problems on heredity and genetics. Here is some information you will need to be able to solve the problems. Please read the information carefully before you begin.

Genotype

1. A genotype (such as for hair shape) is the information that lies in the genes and consists of two alleles (e.g., BB, Bb, or bb).

Phenotype

2. A phenotype is how this information is reflected in physical characteristics (e.g., curly or straight hair).

Dominant and Recessive Alleles

3. A dominant allele (represented by a capital/uppercase letter) produces the dominant phenotype. If individuals have even only one dominant allele in their genotype, they will have the dominant phenotype. This dominant allele can come from one parent or both.
4. For a recessive allele (represented by a lowercase letter), to produce a recessive phenotype, the individual's genotype must have two copies of the recessive allele, which means they received one recessive allele from each parent.
5. For hair shape, the dominant allele for curly hair is (H) and the recessive allele for straight hair is (h).
6. Each genotype is composed of two alleles (i.e., HH, Hh, or hh). A dominant phenotype (i.e., curly hair) appears when an individual's genotype has at least one copy of the dominant allele (i.e., HH or Hh).
7. A recessive phenotype (i.e., straight hair) only appears when an individual's genotype has two recessive alleles (i.e., hh).

Homozygous or heterozygous?

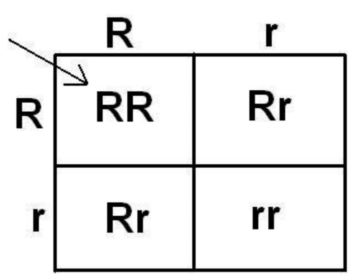
8. Homozygous means that both alleles are the same. This could mean there are two uppercase letters or two lowercase letters (i.e., HH or hh).
9. Heterozygous means that there is a dominant and a recessive allele in the genotype (i.e., Hh).

Type of Reasoning

10. Deductive reasoning means you reason from the parents down to the child.
11. Inductive reasoning means you reason from a child and one parent up to the other parent.

Cross Tables/Punnett Squares

12. To use a cross table/Punnett square, put the two alleles from the genotype of the first parent on the left-hand side of the cross table so that one allele is in each row (e.g., see image below). For the second parent, put one allele on top of the first column and the second allele in the next column. In this example, each parent is heterozygous for a trait. Combine each of the alleles, one from the row and one from the column, in the appropriate boxes to form all the possible genotypes of their child. It is then possible that their child could have the genotypes of RR, Rr, OR rr for this trait.



Appendix D

Appendix D.1. Level 1 Complexity Hereditary Example Problem

Eye color

Eye color in humans is partly determined by a gene, with brown (B) being dominant and blue (b) recessive. Two parents are having a child. Both parents have brown eyes and are homozygous for this trait.

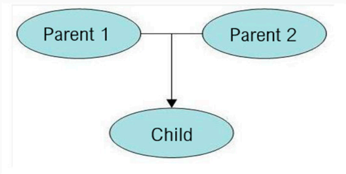
What would be the possible genotype(s) of the child?

Step 1: Translate into genotypes *(1 point)*

Parent 1

Parent 2

Step 2: Enter the genotypes from step 1 in a family chart. Please use the boxes below the example. For the unknown family member(s) you can use a question mark. *(1 point)*



Step 3: Determine direction of your reasoning. *(1 point)*

- ◯ Deductive
- ◯ Inductive

Step 4: Determine the number of cross tables you need and fill them in.

- ◯ 1 cross table
- ◯ 2 cross tables
- ◯ 3 cross tables

Step 5: Answer *(1 point)*

Compare the outcomes you received from the cross table(s) and use them to determine your final answer. Mark the right answer below.

What would be the possible genotype(s) of the child?

- ◯ BB
- ◯ Bb
- ◯ bb
- ◯ BB and Bb
- ◯ BB and bb
- ◯ Bb and bb
- ◯ BB, Bb, and bb

Appendix D.2. Level 3 Complexity Hereditary Example Problem

Eye Color

Eye colour in humans is partly determined by a gene, with brown (B) being dominant and blue (b) recessive. Two parents are having a child. One parent has brown eyes and is heterozygous for this trait. No information is known about the other parent. The child they are having has blue eyes and is homozygous for this trait.

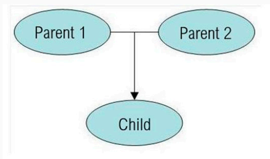
What could be the possible genotype(s) of the parent?

Step 1: Translate to genotypes *(1 point)*

Parent 1

Child

Step 2: Enter the genotypes from step 1 in a family chart. Please use the boxes below the example. For the unknown family member(s) you can use a question mark. *(1 point)*



Step 3: Determine direction of your reasoning. *(1 point)*

- ○ Deductive
- ○ Inductive

Step 4: Determine the number of cross tables you need and fill them in.

- ○ 1 cross table
- ○ 2 cross tables
- ○ 3 cross tables

Cross table 1

		Information parent 2	Information parent 2
Information parent 1			
Information parent 1			

Step 5: Answer *(1 point)*

Compare the outcomes you received from the cross table(s) and use them to determine your final answer. Mark the right answer below.

What could be the possible genotypes of the parent?

- ○ BB
- ○ Bb
- ○ bb
- ○ BB and Bb
- ○ BB and bb
- ○ Bb and bb
- ○ BB, Bb and bb

Appendix D.3. Level 5 Complexity Hereditary Example Problem

Hair type

Hair type in humans is partly determined by a gene, with curly hair (H) being dominant and straight hair (h) being recessive. Two parents (generation P) have a child (F1). One parent has curly hair and is homozygous for this trait. The other parent has straight hair and is homozygous for this trait as well. No information is known about their child. This child (F1) and their partner have a (grand)child (F2). The partner has curly hair and is heterozygous for this trait. No information is known about the (grand)child (F2).

What could be the possible genotype(s) of the child (F1) and the (grand)child (F2)?

Step 1: Translate to genotypes *(1 point)*

Parent 1

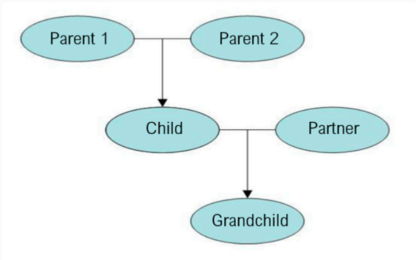
Parent 2

Child

Partner

Grandchild

Step 2: Enter the genotypes from step 1 in to the family chart. Please use the boxes below the example. For unknown family members you can use a question mark. *(1 point)*



Step 3: You will solve this question in two steps. For the first set of cross tables, determine the direction of your reasoning. *(1 point)*

- ◯ Deductive
- ◯ Inductive

Step 4: Determine the number of cross tables you need and fill them in.

- ◯ 1 cross table
- ◯ 2 cross tables
- ◯ 3 cross tables

Step 5: Determine the direction of your reasoning for the second set of cross tables. *(1 point)*

- ◯ Deductive
- ◯ Inductive

Step 6: Determine the number of cross tables you need and fill them in.

- ◯ 1 cross table
- ◯ 2 cross tables
- ◯ 3 cross tables
- ◯ 4 cross tables
- ◯ 5 cross tables
- ◯ 6 cross tables

Step 7: Unknown genotype of the child (F1) *(1 point)*

Compare the outcomes you received from the cross tables and use them to determine your final answer. Mark the right answer below.

- ◯ HH
- ◯ Hh
- ◯ hh
- ◯ HH and Hh
- ◯ HH and hh
- ◯ Hh and hh
- ◯ HH, Hh and hh

Step 8: Unknown genotype of the grandchild (F1) *(1 point)*

Compare the outcomes you received from the cross tables and use them to determine your final answer. Mark the right answer below.

- ◯ HH
- ◯ Hh
- ◯ hh
- ◯ HH and Hh
- ◯ HH and hh
- ◯ Hh and hh
- ◯ HH, Hh and hh

Appendix E.

Mental Effort Rating Scale Adapted from Paas [32]

How much mental effort did you invest to solve this problem?

	very, very low mental effort 1	2	low mental effort 3	4	not high and not low mental effort 5	6	high mental effort 7	8	very, very high mental effort 9
Mental effort	◯	◯	◯	◯	◯	◯	◯	◯	◯

Appendix F

Appendix F.1. Complexity Levels 1–3 Self-Assessment Measure

You can get a maximum of 4 points for solving the previous problem. How many points do you think you scored?

	0 points	1 point	2 points	3 points	4 points
Points Scored	○	○	○	○	○

Appendix F.2. Complexity Level 4 Self-Assessment Measure

You can get a maximum of 5 points for solving the previous problem. How many points do you think you scored?

	0 points	1 point	2 points	3 points	4 points	5 points
Points Scored	○	○	○	○	○	○

Appendix F.3. Complexity Level 5 Self-Assessment Measure

You can get a maximum of 6 points for solving the previous problem. How many points do you think you scored?

	0 points	1 point	2 points	3 points	4 points	5 points	6 points
Points Scored	○	○	○	○	○	○	○

Appendix G.

Appendix G.1. Instructions for the Problem-Solving Tasks for the Control Group

In the next section, you will be solving 10 problems. When working through each of these problems, please **DO** the following:

1. Try to solve the questions as quickly and accurately as you can.
2. If you absolutely cannot answer a question, please put a question mark (?) as the answer.

Do **NOT**:

1. Do ***NOT*** use your hands, write things down, or speak when solving the problems.

Please click yes if you understand and accept the instructions.

○ Yes

Appendix G.2. Instructions for the Problem-Solving Tasks for the Experimental Group

In the next section, you will be solving 10 problems. When working through each of these problems, please **DO** the following:

1. ***Think aloud*** (say what you are doing and thinking as you are doing or thinking it) as you work through each of the problems. Please do this ***in English***.
2. Try to solve the questions as quickly and accurately as you can.
3. If you absolutely cannot answer a question, please put a question mark (?) as the answer.

Do ***NOT:***

1. Do ***NOT*** write anything down or use your hands when solving the problems.
2. Do ***NOT*** say any personally identifiable information such as your name or age during the recordings.

If you do not want your voice recorded, you may stop your participation now or later without penalty or justification. Also, a reminder you have the right to ask for your data not to be used.

The audio recording will start automatically, so just follow the think aloud instructions as soon as you start the problems!

Please click yes if you understand and accept the instructions.

○ Yes

→

Appendix H.

Coding Scheme for Qualitative Analysis (Based on Azevedo et al. [45])		
Variable	Description	Example
A—Forethought, Planning, and Activation		
A1—Planning	Includes listing the actions needed to solve the problem before executing them.	*"I need to do the table first"*
A2—Sub-Goals	Participant states actions that are intended to help solve the problem while solving the problem.	*"And then so we know that so we can do deductive again"*
A3—Prior Knowledge Activation	Participant recalls relevant prior knowledge before a task or during.	*"Homozygous means they are both the same gene"*
A4—Recycle Goal in Working Memory	Participant repeats the goal verbally to refresh it in their working memory.	*"Both parents have brown eyes. . .both parents have brown eyes"*

B—Monitoring		
B1—Negative Judgment of Learning (NJOL)	Participant becomes aware that they do not understand or know the material during problem-solving.	*"Still not sure"* *"The child, wait hold on, this is so confusing, what am I doing?"*
B2—Positive Judgment of Learning (PJOL)	Participant becomes aware that they do understand or know the material during problem solving.	*"Which means yeah that makes sense"* *"So parent one is big B, big B. Right? Yeah"*
B3—Self-assessment	Participate gives a positive or negative self-assessment of their performance (when prompted to during self-assessment rating question)	*"I think I scored all of them, probably"* *"Guessing four points hopefully"*
B4—Feeling of Knowing (FOK)	Participant states that they remember the material vaguely and understand some of it, but not well.	*"There was another way to do this, but I forgot"* *"And it can either be both recessive or large, I think that's possible? Yeah"*
B5—Content Evaluation	Participant monitors content relative to goals.	*"I don't think the child's eye colour was mentioned even"* *"So the parent has a big B, otherwise they would not had brown eyes"*
B6—Identify Adequacy of Information and Task Instructions	Participant determines the usefulness or adequacy of the material and task instructions.	*"So I am just going to put information parent two as... I don't know how to put that in there, honestly"* *"I think we are talking about the child here, because that is the only one that is kinda missing and I would technically say it is both...oh for the first set of the cross tables, okay okay that makes a lot of sense now."*
C—Strategy Use		
C1—Goal-Free Search	Participant states that they will look around the information without stating a particular goal or plan.	[Did not appear in any the transcripts]
C2—Summarization	Participant summarizes what they read.	*"The partner has curly hair and is heterozygotic, so they have the same traits as the parents of the child"* *"Okay, so one parent is homozygous and dominant and the other one is heterozygous"*
C3—Rereading	Participants reread or revisit earlier material.	*"And then the partner had heterozygotic set, right? Curly hair, yes. And is heterozygotic."*
C4—Inferences	Participant makes inferences based on what they read or saw in the environment.	*"So both have brown eyes and are homozygous so big B big B."* *"Deductive reasoning"* *"Which means the parent can be big little and then little. So then this one"*
C5—Hypothesizing	Participant asks questions that go beyond what they read in the material.	*"But these are the options that the child definitely does not have, if it is two small h's it would not work. But it is probabilities so who knows."* *"So we have a 50/50 chance that the child will have homozygous brown eyes or homozygous... heterozygous brown eyes."*
C6—Knowledge Elaboration	Participant elaborates what they read in the material with prior knowledge.	*"Ok so h h. And to have a child that has one big H that means that they have to have at least one big h."* *"Again this is deductive, because we are looking at the parents and then trying to guess the child one"*
D—Task Difficulty, Complexity, and Demands		
D1—Time and Effort Planning	Participant deliberately controls their behavior due to limited energy or other physical demands.	*"Just going to say small instead of lowercase or uppercase"* *"Since question is going to ask I'm just going to fill the first one because I can't"*

D2—Help-Seeking Behavior	Participant seeks assistance from the experimenter regarding the task.	[Any dialogues with the experimenters about the task]
D3—Task Difficulty	Participant indicates that the task is easy or difficult.	*"It's a brawl of kk, this is easy"* *"Okay, this is a bit more complex"*
D4—Mental Effort	Participant gives an indication of their mental effort invested in the task.	*"I guess same mental effort or a bit more maybe"* *"Say low, probably. But I did think about it"*
E—Miscellaneous Actions		
E1—Stating Answer	Participant states an answer to a final step in the problem-solving task.	*"All the possibilities"*
E2—Copying information	Participant copies information they wrote again or found in the environment or verbalizes information they input into a blank or cross table.	*"Parent one is big big, parent two is little little"* *"So big big little big big and big little"*
E3—Reading Question	Participant reads questions or information (close to) verbatim on screen out loud.	*"Two parents are having a child."*

Note. All italicized examples are direct quotes from the transcripts. All unitalicized examples in brackets are explanations

List of Adjustments from Base Coding Scheme

- Simplified or adjusted definition for this context and changed example to relevant example from the transcripts
    - A1—Planning
    - A3—Prior Knowledge Activation
    - A4—Recycle Goal in Working Memory
    - B4—Feeling of Knowing (FOK)
    - C2—Summarization
    - C3—Re-reading
    - C4—Inferences
    - C5—Hypothesizing
    - C6—Knowledge Elaboration
    - D1—Time and Effort Planning
    - D2—Help-Seeking Behavior
    - D3—Task Difficulty
- Heavily adjusted definition for this context, moved it to a completely new code group, and put relevant example taken from the transcripts
    - E2—Copying information
- Simplified definition for this context and gave a hypothetical example not from the transcripts
    - A2—Sub-Goals
    - C1—Goal-Free Search
- Definition kept the same but relevant example from the transcripts replaced original example
    - B5—Content Evaluation
- Added this code due to frequent occurrences in the text and/or relevancy to research
    - E3—Reading Question
    - D4—Mental Effort
    - E1—Stating Answer
- Originally one code of called judgments of learning, this code was split up into two different codes, opposites of each other, for more in depth analysis of kinds of judgments of learning the participants verbalized. Definitions were adjusted and relevant examples from the transcripts were given.
    - B1—Negative Judgment of Learning (NJOL)
    - B2—Positive Judgment of Learning (PJOL)

- This code was added to the monitoring code group to be able to separate out when participants monitored spontaneously (such as a JOL) and when participants monitored their performance when prompted in the task environment when doing the self-assessment rating scale.
    - B3—Self-assessment
- Originally the two codes of "Identify Adequacy of Information" and "Expectation of Adequacy of Information", these codes were combined together to form this code, which included judgments based on the usefulness and adequacy of task instructions and material presented, which fit better with this context. Definition was adjusted to fit the new code and relevant examples from the transcripts were added.
    - B6—Identify Adequacy of Information and Task Instructions
- These codes were removed due to irrelevance or redundancy.
    - Self-Questioning
    - Selecting a New Informational Source
    - Control of Context
- Code group called Task Difficulty and Demands was renamed to Task Difficulty, Complexity, and Demands
- Added code group called Miscellaneous Actions due to these codes not fitting in with other code groups well but being frequent types of verbalizations

Appendix I.

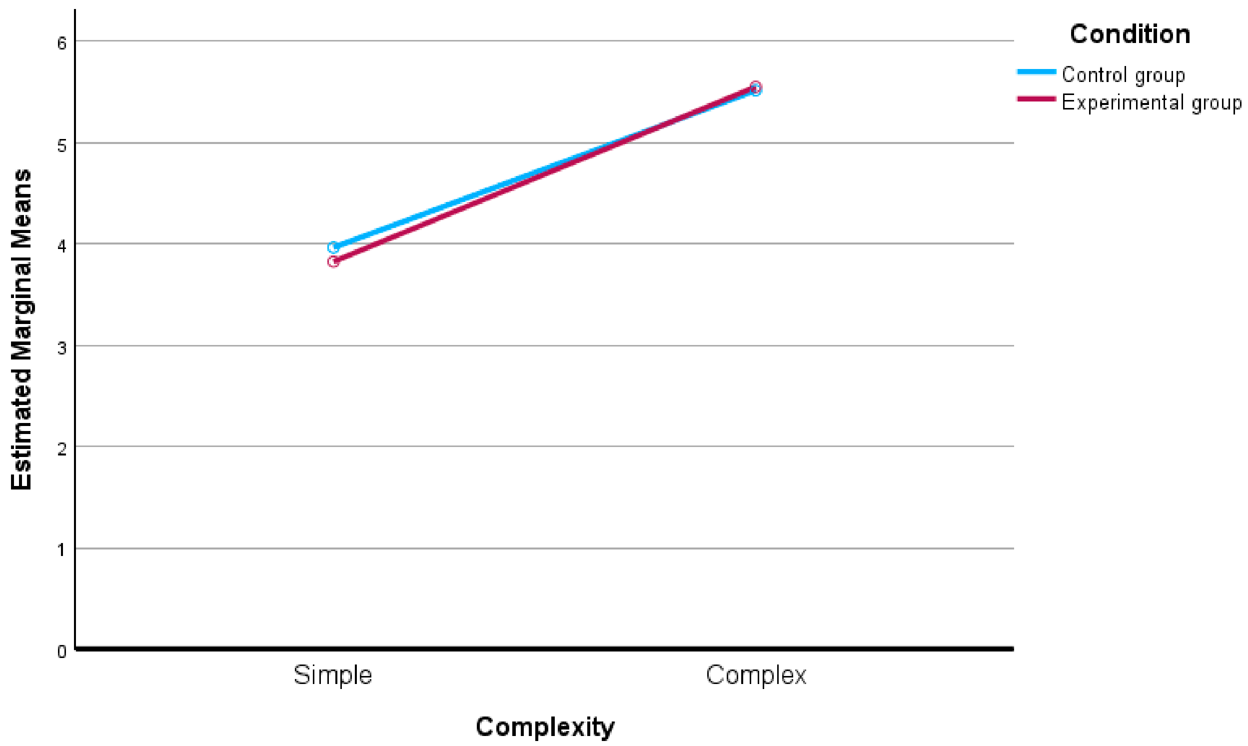


Figure A1. Average mental effort rating per condition for both complexity levels.

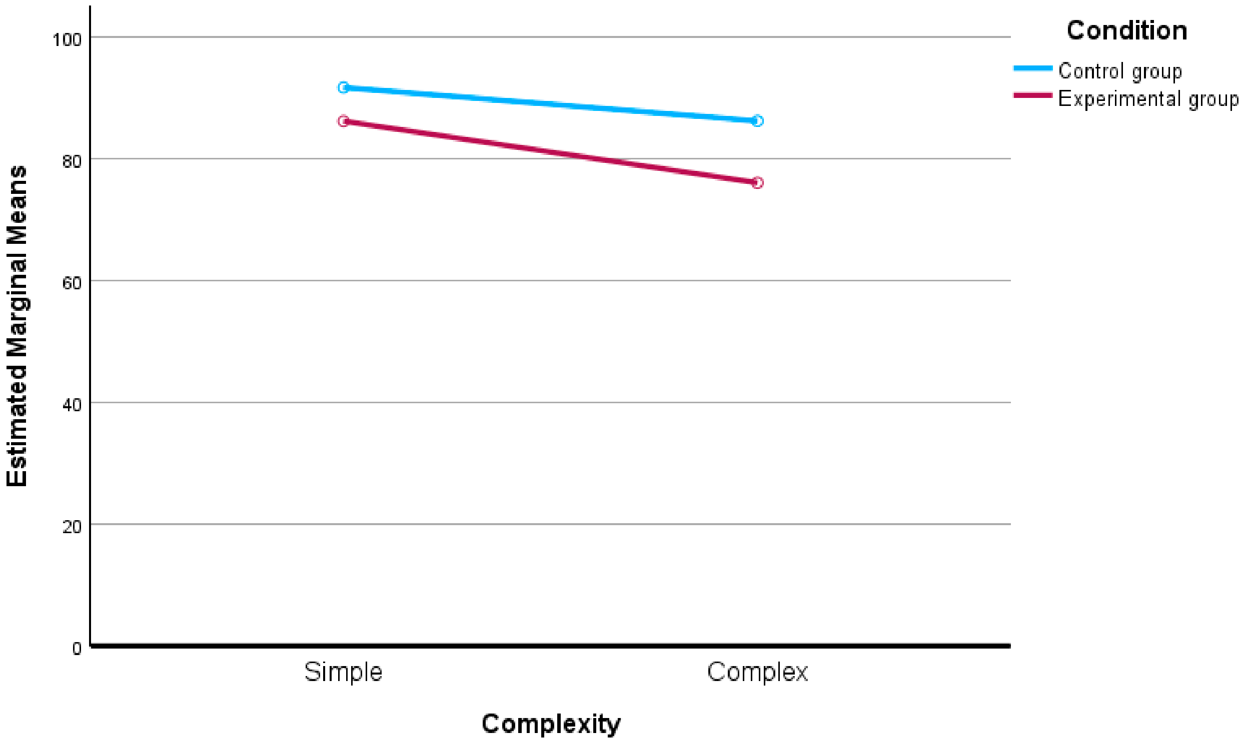


Figure A2. Average performance score per condition for both complexity levels.

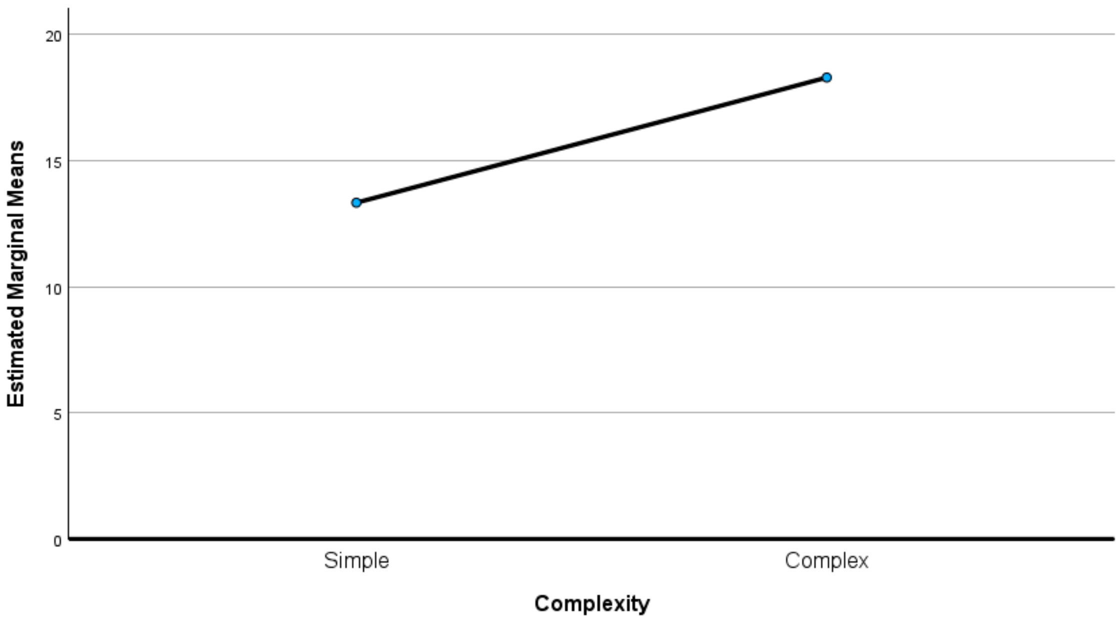


Figure A3. Average absolute self-assessment accuracy score of the experimental group for both complexity levels.

Appendix J.

Pairwise Comparisons for Average Performance per Complexity Level

Complexity (I)	Complexity (J)	Mean Difference (I–J)	Standard Error	Sig.	95% CI for Difference Lower Bound	Upper Bound
1	2	1.310	2.299	1.000	−5.380	7.999
	3	7.530	2.572	0.047	0.047	15.012
	4	9.292	2.757	0.013	1.272	17.311
	5	8.562	2.520	0.012	1.230	15.893
2	1	−1.310	2.299	1.000	−7.999	5.380
	3	6.220	1.919	0.019	0.637	11.803
	4	7.982	2.264	0.008	1.396	14.568
	5	7.252	2.106	0.010	1.126	13.378
3	1	−7.530	2.572	0.047	−15.012	−0.047
	2	−6.220	1.919	0.019	−11.803	−0.637
	4	1.762	2.120	1.000	−4.407	7.930
	5	1.032	1.983	1.000	−4.739	6.802
4	1	−9.292	2.757	0.013	−17.311	−1.272
	2	−7.982	2.264	0.008	−14.568	−1.396
	3	−1.762	2.120	1.000	−7.930	4.407
	5	−0.730	1.994	1.000	−6.531	5.071
5	1	−8.562	2.520	0.012	−15.893	−1.230
	2	−7.252	2.106	0.010	−13.378	−1.126
	3	−1.032	1.983	1.000	−6.802	4.739
	4	0.730	1.994	1.000	−5.071	6.531

Note: For the performance measure, complexity levels one and two are significantly different from levels three through five, but they are similar to each other. Meanwhile, complexity levels three through five also are statistically similar to each other. Therefore, for the greater analysis, we decided to average together the performances on levels one and two for a simple performance mean and then averaged levels three through five together to obtain a complex performance mean. This was similarly done with mental effort to obtain a simple and complex mental effort mean and absolute SA accuracy to obtain a simple and complex absolute SA accuracy mean.

References

1. Bjork, R.A.; Dunlosky, J.; Kornell, N. Self-regulated learning: Beliefs, techniques, and illusions. *Annu. Rev. Psychol.* **2013**, *64*, 417–444. [CrossRef] [PubMed]
2. Zimmerman, B.J. Becoming a self-regulated learner: Which are the key subprocesses? *Contemp. Educ. Psychol.* **1986**, *11*, 307–313. [CrossRef]
3. Stefanou, C.; Stolk, J.D.; Prince, M.; Chen, J.C.; Lord, S.M. Self-regulation and autonomy in problem- and project-based learning environments. *Act. Learn. High. Educ.* **2013**, *14*, 109–122. [CrossRef]
4. Lai, C.; Hwang, G. A self-regulated flipped classroom approach to improving students' learning performance in a mathematics course. *Comput. Educ.* **2016**, *100*, 126–140. [CrossRef]
5. Littlejohn, A.; Hood, N.; Milligan, C.; Mustain, P. Learning in MOOCs: Motivations and self-regulated learning in MOOCs. *Internet High. Educ.* **2016**, *29*, 40–48. [CrossRef]
6. Winne, P.H. A metacognitive view of individual differences in self-regulated learning. *Learn. Individ. Differ.* **1996**, *8*, 327–353. [CrossRef]
7. Azevedo, R.; Cromley, J. Does training on self-regulated learning facilitate students' learning with hypermedia? *J. Educ. Psychol.* **2004**, *96*, 523–535. [CrossRef]
8. Baars, M.; Visser, S.; Van Gog, T.; de Bruin, A.; Paas, F. Completion of partially worked-out examples as a generation strategy for improving monitoring accuracy. *Contemp. Educ. Psychol.* **2013**, *38*, 395–406. [CrossRef]
9. Baars, M.; Van Gog, T.; de Bruin, A.; Paas, F. Effects of problem solving after worked example study on primary school children's monitoring accuracy. *Appl. Cogn. Psychol.* **2014**, *28*, 382–391. [CrossRef]
10. Baars, M.; Vink, S.; Van Gog, T.; De Bruin, A.; Paas, F. Effects of training self-assessment and using assessment standards on retrospective and prospective monitoring of problem solving. *Learn. Instr.* **2014**, *33*, 92–107. [CrossRef]

11. Van Gog, T.; Kester, L.; Paas, F. Effects of concurrent monitoring on cognitive load and performance as a function of task complexity. *Appl. Cogn. Psychol.* **2011**, *25*, 584–587. [CrossRef]
12. De Bruin, A.B.; Roelle, J.; Carpenter, S.K.; Baars, M.; EFG-MRE. Synthesizing cognitive load and self-regulation theory: A theoretical framework and research agenda. *Educ. Psychol. Rev.* **2020**, *32*, 903–915. [CrossRef]
13. Sweller, J.; Van Merrienboer, J.J.G.; Paas, F.G.W.C. Cognitive architecture and instructional design. *Educ. Psychol. Rev.* **1998**, *10*, 251–296. [CrossRef]
14. Sweller, J.; Van Merrienboer, J.J.G.; Paas, F.G.W.C. Cognitive architecture and instructional design: 20 years later. *Educ. Psychol. Rev.* **2019**, *31*, 261–292. [CrossRef]
15. Baddeley, A. Working memory. *Science* **1992**, *255*, 556–559. [CrossRef]
16. Paas, F.G.W.C.; Van Merrienboer, J.J.G. Instructional-control of cognitive load in the training of complex cognitive tasks. *Educ. Psychol. Rev.* **1994**, *6*, 351–371. [CrossRef]
17. Brünken, R.; Plass, J.L.; Leutner, D. Direct measurement of cognitive load in multimedia learning. *Educ. Psychol.* **2003**, *38*, 53–61. [CrossRef]
18. Paas, F.; Van Gog, T.; Sweller, J. Cognitive load theory: New conceptualizations, specifications, and integrated research perspectives. *Educ. Psychol. Rev.* **2010**, *22*, 115–121. [CrossRef]
19. Radović, S.; Seidel, N.; Haake, J.M.; Kasakowskij, R. Analyzing students' self-assessment practice in a distance education environment: Student behavior, accuracy, and task-related characteristics. *J. Comput. Assist. Learn.* **2023**, *40*, 654–666. [CrossRef]
20. Pintrich, P.R. The role of goal orientation in self-regulated learning. In *Handbook of Self-Regulation*; Boekaerts, M., Pintrich, P.R., Zeidner, M., Eds.; Academic Press: Cambridge, MA, USA, 2000; pp. 451–502. [CrossRef]
21. Winne, P.H.; Hadwin, A.F. Studying as self-regulated learning. In *Metacognition in Educational Theory and Practice*, 2nd ed.; Hacker, D., Dunlosky, J., Graesser, A., Eds.; Lawrence Erlbaum: Mahwah, NJ, USA, 1998; pp. 277–304. [CrossRef]
22. Panadero, E. A review of self-regulated learning: Six models and four directions for research. *Front. Psychol.* **2017**, *8*, 422. [CrossRef]
23. Winne, P.H. Cognition and metacognition within self-regulated learning. In *Handbook of Self-Regulation of Learning and Performance*, 2nd ed.; Schunk, D.H., Greene, J.A., Eds.; Routledge: London, UK, 2017; pp. 36–48. [CrossRef]
24. Ku, K.Y.L.; Ho, I.T. Metacognitive strategies that enhance critical thinking. *Metacogn. Learn.* **2010**, *5*, 251–267. [CrossRef]
25. Mirandola, C.; Ciriello, A.; Gigli, M.; Cornoldi, C. Metacognitive monitoring of text comprehension: An investigation on postdictive judgments in typically developing children and children with reading comprehension difficulties. *Front. Psychol.* **2018**, *9*, 2253. [CrossRef] [PubMed]
26. Barenberg, J.; Dutke, S. Testing and metacognition: Retrieval practise effects on metacognitive monitoring in learning from text. *Memory* **2019**, *27*, 269–279. [CrossRef] [PubMed]
27. Dunlosky, J.; Rawson, K.A. Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learn. Instr.* **2012**, *22*, 271–280. [CrossRef]
28. Pieschl, S. Metacognitive calibration—An extended conceptualization and potential applications. *Metacogn. Learn.* **2009**, *4*, 3–31. [CrossRef]
29. Peng, Y.; Tullis, J.G. Dividing attention impairs metacognitive control more than monitoring. *Psychon. Bull. Rev.* **2021**, *28*, 2064–2074. [CrossRef]
30. Tuysuzoglu, B.B.; Greene, J.A. An investigation of the role of contingent metacognitive behavior in self-regulated learning. *Metacogn. Learn.* **2015**, *10*, 77–98. [CrossRef]
31. Seufert, T. The interplay between self-regulation in learning and cognitive load. *Educ. Res. Rev.* **2018**, *24*, 116–129. [CrossRef]
32. Paas, F.G.W.C. Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *J. Educ. Psychol.* **1992**, *84*, 429–434. [CrossRef]
33. Winne, P.H. Inherent details in self-regulated learning. *Educ. Psychol.* **1995**, *30*, 173–187. [CrossRef]
34. Vangsness, L.; Young, M.E. More isn't always better: When metacognitive prompts are misleading. *Metacogn. Learn.* **2021**, *16*, 135–156. [CrossRef]
35. Liefooghe, B.; Barrouillet, P.; Vandierendonck, A.; Camos, V. Working memory costs of task switching. *J. Exp. Psychol. Learn. Mem. Cogn.* **2008**, *34*, 478–494. [CrossRef] [PubMed]
36. Patterson, M.C. A naturalistic investigation of media multitasking while studying and the effects on exam performance. *Teach. Psychol.* **2017**, *44*, 51–57. [CrossRef]
37. Finley, J.R.; Benjamin, A.S.; McCarley, J.S. Metacognition of multitasking: How well do we predict the costs of divided attention. *J. Exp. Psychol. Appl.* **2014**, *20*, 158–165. [CrossRef]
38. Trabasso, T.; Magliano, J.P. Conscious understanding during comprehension. *Discourse Process.* **1996**, *21*, 255–287. [CrossRef]
39. Meyers, J.; Lytle, S.; Palladino, D.; Devenpeck, G.; Green, M. Think-aloud protocol analysis: An investigation of reading comprehension strategies in fourth- and fifth-grade students. *J. Psychoeduc. Assess.* **1990**, *8*, 112–127. [CrossRef]
40. Schellings, G.L.M.; Van Hout-Wolters, B.H.A.M.; Veenman, M.V.J.; Meijer, J. Assessing metacognitive activities: The in-depth comparison of a task-specific questionnaire with think-aloud protocols. *Eur. J. Psychol. Educ.* **2013**, *28*, 963–990. [CrossRef]
41. Zachariou, A.; Whitebread, D. Developmental differences in young children's self-regulation. *J. Appl. Dev. Psychol.* **2019**, *62*, 282–293. [CrossRef]

42. Jordano, M.L.; Touron, D.R. How often are thoughts metacognitive? Findings from research on self-regulated learning, think-aloud protocols, and mind-wandering. *Psychon. Bull. Rev.* **2018**, *25*, 1269–1286. [CrossRef]
43. Qualtrics XM—Experience Management Software. Available online: https://www.qualtrics.com/ (accessed on 5 May 2023).
44. Moos, D.C.; Bonde, C. Flipping the classroom: Embedding self-regulated learning prompts in videos. *Technol. Knowl. Learn.* **2016**, *21*, 225–242. [CrossRef]
45. Azevedo, R.; Guthrie, J.T.; Seibert, D. The role of self-regulated learning in fostering students' conceptual understanding of complex systems with hypermedia. *J. Educ. Comput. Res.* **2004**, *30*, 87–111. [CrossRef]
46. Johnstone, C.J.; Bottsford-Miller, N.A.; Thompson, S.J.; Council of Chief State School Officers (CCSSO); National Association of State Directors of Special Education (NASDSE). *Using the Think Aloud Method (Cognitive Labs) to Evaluate Test Design for Students with Disabilities and English Language Learners*; University of Minnesota, National Center on Educational Outcomes: Minneapolis, MN, USA, 2006.
47. ATLAS.ti Scientific Software Development GmbH. ATLAS.ti Windows (Version 24.0.0) [Qualitative Data Analysis Software]. 2024. Available online: https://atlasti.com (accessed on 1 January 2024).
48. IBM Corp. *IBM SPSS Statistics for Windows, version 29.0*; Computer software; IBM Corp.: Armonk, NY, USA, 2023.
49. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed.; Lawrence Erlbaum Associates: Mahwah, NJ, USA, 1988.
50. Leppink, J.; Paas, F.; Van der Vleuten, C.P.M.; Van Gog, T.; Van Merriënboer, J.J.G. Development of an instrument for measuring different types of cognitive load. *Behav. Res. Methods* **2013**, *45*, 1058–1072. [CrossRef] [PubMed]
51. Hoch, E.; Sidi, Y.; Ackerman, R.; Hoogerheide, V.; Scheiter, K. Comparing mental effort, difficulty, and confidence appraisals in problem-solving: A metacognitive perspective. *Educ. Psychol. Rev.* **2023**, *35*, 61. [CrossRef]
52. Albulescu, P.; Macsinga, I.; Rusu, A.; Sulea, C.; Bodnaru, A.; Tulbure, B.T. "Give me a break!" A systematic review and meta-analysis on the efficacy of micro-breaks for increasing well-being and performance. *PLoS ONE* **2022**, *17*, e0272460. [CrossRef] [PubMed]
53. Tyler, J.M.; Burns, K.C. After depletion: The replenishment of the self's regulatory resources. *Self Identity* **2008**, *7*, 305–321. [CrossRef]
54. Steinborn, M.B.; Huestegge, L. A walk down the lane gives wings to your brain. Restorative benefits of rest breaks on cognition and self-control. *Appl. Cogn. Psychol.* **2016**, *30*, 795–805. [CrossRef]
55. Nietfeld, J.L.; Cao, L.; Osborne, J.W. Metacognitive monitoring accuracy and student performance in the postsecondary classroom. *J. Exp. Educ.* **2005**, *74*, 7–28.